Testing for mildly versus strongly misspecified models

Anonymous Author(s)
Affiliation
Address
email

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Included as a link to github
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Everything is specified (and all details are given in the code on github)
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The simualted experiments were repeated 1000 times and the histograms of p-values are presented.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] The time to compute is of order of seconds.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Derivation of D and V for multinomial model

We consider the following probability model. Given counts of words $n_k^{(\ell)}$, $\sum_{k=1}^p n_k^{(\ell)} = n^{(\ell)}$ for separate (independent) texts $\ell = 1, \ldots, N$ by the same author, assume that the corresponding probabilities are the same in all texts given by θ_k, $\sum_{k=1}^p \theta_k = 1$. So here we have $N = 11$ books, with $n^{(\ell)}$ words in each book. The multinomial model for such that has the following likelihood:

$$L(\theta; (n_k^{(\ell)})) = \prod_{\ell=1}^{N} \prod_{k=1}^{p} \theta_k^{n_k^{(\ell)}}.$$

Only $p-1$ unknown parameters are independent. Then,

$$\ell(\theta) = \sum_{\ell=1}^{N} \sum_{k=1}^{p-1} n_k^{(\ell)} \log \theta_k + n_p^{(\ell)} \log \left(1 - \sum_{k=1}^{p-1} \theta_k\right)$$

and for $j = 1, 2, \ldots, p-1$, denoting $N_j = \sum_{\ell=1}^N n_j^{(\ell)}$:

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_j} &= N_j/\theta_j - N_p/\left(1 - \sum_{k=1}^{p-1} \theta_k\right) \\
\frac{\partial^2 \ell(\theta)}{\partial \theta_j^2} &= -N_j/\theta_j^2 - N_p/\left(1 - \sum_{k=1}^{p-1} \theta_k\right)^2 \\
\frac{\partial^2 \ell(\theta)}{\partial \theta_j \partial \theta_m} &= -N_p/\left(1 - \sum_{k=1}^{p-1} \theta_k\right)^2
\end{aligned}$$

so, using $Cov(n_j^{(\ell)}, n_k^{(\ell)}) = -n^{(\ell)}\theta_j\theta_k$, we have

$$\begin{aligned}
V_{j,m}(\theta) &= E\left(\frac{\partial \ell(\theta)}{\partial \theta_j}\frac{\partial \ell(\theta)}{\partial \theta_m}\right) \\
&= \mathbb{E}\left(N_j/\theta_j - N_p/\left(1 - \sum_{k=1}^{p-1} \theta_k\right)\right)\left(N_m/\theta_m - N_p/\left(1 - \sum_{k=1}^{p-1} \theta_k\right)\right) \\
&= \sum_{\ell=1}^{N} [Cov(n_j^{(\ell)}/\theta_j, n_m^{(\ell)}/\theta_m) + Var(n_p^{(\ell)}/\theta_p) - Cov(n_p^{(\ell)}/\theta_p, n_j^{(\ell)}/\theta_j) - Cov(n_p^{(\ell)}/\theta_p, n_m^{(\ell)}/\theta_m)]
\end{aligned}$$

with $\theta_p = 1 - \sum_{k=1}^{p-1} \theta_k$, and on the diagonal

$$\begin{aligned}
V_{j,j}(\theta) &= E\left(\frac{\partial \ell(\theta)}{\partial \theta_j}\right)^2 \\
&= \sum_{\ell=1}^{N} \mathbb{E}\left(n_j^{(\ell)}/\theta_j - n_p^{(\ell)}/\theta_p\right)^2 \\
&= \sum_{\ell=1}^{N} \left[Var(n_j^{(\ell)})/\theta_j^2 + Var(n_p^{(\ell)})/\theta_p^2 + -2Cov(n_j^{(\ell)}, n_p^{(\ell)})/[\theta_j\theta_p]\right].
\end{aligned}$$

Also,

$$\begin{aligned}
D_{jj}(\theta) &= \mathbb{E}N_j/\theta_j^2 + \mathbb{E}N_p/\theta_p^2 = \sum_{\ell=1}^{N} n^{(\ell)}[1/\theta_j + 1/\theta_p], \\
D_{jm}(\theta) &= \mathbb{E}N_p/\theta_p^2 = \sum_{\ell=1}^{N} n^{(\ell)}/\theta_p].
\end{aligned}$$

The (p)MLE is

$$\hat{\theta}_j = N_j/[\sum_{\ell=1}^{N} n^{(\ell)}], \quad j = 1, \dots, p.$$

Denote $X_j^{(\ell)} = n_j^{(\ell)}/\hat{\theta}_j - n^{(\ell)}$, then

$$\begin{aligned} \hat{V}_{j,m} &= V_{y,jm}(\hat{\theta}) = \sum_{\ell=1}^{N} \left[X_j^{(\ell)} X_m^{(\ell)} + [X_p^{(\ell)}]^2 + - X_j^{(\ell)} X_p^{(\ell)} - X_m^{(\ell)} X_p^{(\ell)} \right], \\ \hat{V}_{j,j} &= V_{y,jj}(\hat{\theta}) = \sum_{\ell=1}^{N} \left[[X_j^{(\ell)}]^2 + [X_p^{(\ell)}]^2 + -2 X_j^{(\ell)} X_p^{(\ell)} \right] \end{aligned}$$

B Statistical analysis of texts in R

The code used to analyse simulated and text data, with all preprocessing details and the list of books by A. Conan Doyle used in the analysis, is available on `https://github.com/nataliabochkina/TestMisspecifiedModel`.